Model ensemble in density-based clustering

24th International Conference on Computational Statistics

Alessandro Casa

Joint work with: L.Scrucca & G.Menardi

Faculty of Economics and Management

Free University of Bozen-Bolzano

alessandro.casa@unibz.it

Bologna, 25th August 2022

Mixture modelling and clustering

- Model-based clustering offers a probabilistic formalization of the clustering problem
- Let $\mathbf{X} = \{x_1, \ldots, x_n\}$, with $x_i \in \mathbb{R}^p$, be the set of observed data. The density of a generic point x is given by

$$f(x; \Theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k)$$

  - $\Theta = (\pi_1, \ldots, \pi_{K-1}, \theta_1, \ldots, \theta_K)$, with $\pi_k > 0$ and $\sum_k \pi_k = 1$
  - Often $f_k(\cdot) = \phi_k(\cdot)$ with $\theta_k = \{\mu_k, \Sigma_k\}$, with parsimony induced by eigen-decomposition $\Sigma_k = \lambda_k A_k D_k A_k^T$
- MLE of Θ is carried out via EM-algorithm and the partition is obtained resorting to the components-clusters correspondence

Model selection in MBC

- Model selection step is essential to choose a model providing a good clustering. Need to choose:
  - number of cluster K
  - parametrizations of Σ_k
  - component densities f_k
  - ...

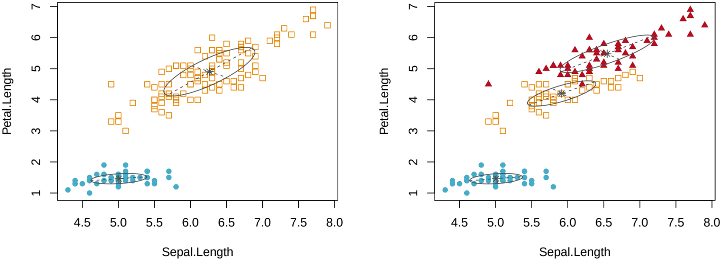
Single-best model paradigm

several different models are fitted, with the best one being chosen according to an IC and used to obtain a partition

- Sub-optimal solution as we are throwing away possibly useful model and neglecting selection-related uncertainty

Super-creative motivating example

- Let see what happens with the `Iris data`



`VEV2, BIC=-561.72` `VEV3, BIC=-562.55`

Aim of the work

propose a model averaging approach in the model-based clustering framework

What has been done before?

- Averaging and ensemble approaches have been rarely studied in the unsupervised framework with respect to the supervised counterpart. Why?
- We need an invariant quantity, having the same meaning for all the models being mixed (not that easy in MBC!)
- Clever workarounds recently proposed:
  - Wei & McNicholas (2015) average *a posteriori* probabilities after a component-merging step
  - Russell et al. (2015) average *similarity matrices*, eventually obtaining partitions via hierarchical clustering

What we propose

Our idea

choose the density as the invariant quantity
thus recasting the problem as a density estimation one

- Resulting density estimate is a convex linear combination of a subset of the fitted models

$$\hat{f}(x;\alpha) = \sum_{m=1}^{M} \alpha_m f_m(x|\hat{\Theta})$$

  where $f_m(\cdot)$ are the models to average, M is their number and α_m the corresponding weights

- Yes, but...
  - How do we estimate α_m's?
  - How do we choose M?
  - How do we practically obtain the partition?

Weights estimation

- $\hat{f}(\cdot)$ is itself a (*simplified*) mixture, so α_m's are estimated maximizing the log-likelihood by means of the EM algorithm
- Overfitting in action, the most complex models (i.e. with more components) will weight more in the combination
- We propose to obtain $\hat{\alpha}$ by maximizing the penalized log-likelihood

$$\ell_P(\alpha|\mathbf{X}) = \sum_{i=1}^{n} \log \sum_{m=1}^{M} \alpha_m f_m(x_i) - \lambda \sum_{m=1}^{M} \alpha_m \nu_m$$

  with ν_m the number of parameters for the m-th model
- (Again) yes, but...
  $\longrightarrow$ how do we select λ?

Penalization strength

- Hyperparameter λ drives the strenght of the penalization thus choosing which models will play a role in the ensemble

Different strategies explored:

- CrossVal-based
  - split iteratively the dataset $\mathbf{X}$ in test and training
  - for λ's in a reasonable grid compute $\hat{f}(x_{\text{test}}|x_{\text{train}})$
  - compute the *test log-likelihood*

$$\ell_{\text{test}} = \sum_{x \in \mathbf{X}_{\text{test}}} \log \hat{f}(x|x_{\text{train}})$$

  and select $\lambda_{\text{CV}} = \text{argmax}\, \ell_{\text{test}}(\lambda)$

- IC-based

  BIC and AIC-type, leading to $\lambda_{\text{BIC}} = \log n/2$ and $\lambda_{\text{AIC}} = 1$

How many models?

- Nowadays in data analysis routines a huge number of models is usually estimated
- We need to choose wisely the *useful* ones, to popolate the ensemble with relevant information

Some strategies

- Subjective selection using prior knowledge
- Build an *Occam's window* using some quantity evaluating the goodness of fit of the candidate models (e.g. BIC)
- Include everything and let the penalization to do the job

Small intermezzo - Density based clustering

- What?
  Link the concept of cluster to features of the density underlying the data
- How?
  - Model-based (parametric)
  - Modal (nonparametric)

Modal clustering: groups corresponding to the domains of attraction of the density modes. It requires:

- a density estimate (usually KDE)
- a method to locale the modal regions (mean-shift, modal EM)

Obtain the partition

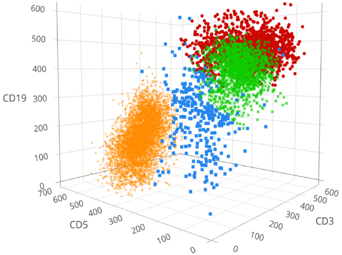
- Our estimate $\hat{f}(x; \hat{\alpha})$ is still a mixture but the correspondence groups-components is lost

Our idea

blend together the two density-based clustering formulations, by shifting the concept of cluster towards the modal one

- We explore the modality of $\hat{f}(x; \hat{\alpha})$ by means of gradient ascent algorithm
  - Modification of the Modal EM algorithm (Scrucca, 2021), an EM-like algorithm searching for local maxima of the density

Some results - Real data



	SB	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.401	0.867	0.909	0.910	0.912
$\hat{K}$	7	4	4	4	4

DLBCL data

- $d = 8$ chemical variables
  $n = 572$ olive oils
- $\hat{K}_{\text{true}} = 9$ regions of Italy
- Hierarchical structure

	SB	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.782	0.792	0.902	0.892	0.902
$\hat{K}$	6	6	8	8	8

Olive Oil data

My two cents

Model-based and modal clustering are
two side of the density-based coin

- Blending them together we reduce weaknesses while mixing respective strenghts
- Two ways to look at the proposal
  - Solution to the single-best model paradigm in the model-based framework
    (+ to rigidity when non-elliptical shapes)
  - Solution to the density estimation problem in the modal framework
- Our estimator lies somewhere in the semi-parametric realm

Remarks, directions & questions

- Overcome the strong reliance of model-based clustering on a single-best model, remaining in a probabilistic framework

Some additional thoughts:

- Post-selection inference: usually ignored in clustering, this approach is more appropriate for uncertainty quantification
- Some contact points with merging and with deep GMM
- What about averaging bootstrap replications or solutions from different initializations?
- Clustering-oriented selection of λ in the penalization strategy

Some references

Casa, A., Scrucca, L., & Menardi, G. (2021).
Better than the best? Answers via model ensemble in density-based clustering. *Advances in Data Analysis and Classification*, 15(3), 599-623.

Other relevant references

- Russell, N., Murphy, T. B., & Raftery, A. E. (2015). Bayesian model averaging in model-based clustering and density estimation. *arXiv:1506.09035.*
- Scrucca, L. (2021). A fast and efficient Modal EM algorithm for Gaussian mixtures. *Statistical Analysis and Data Mining*, 14(4), 305-314.
- Smyth, P., & Wolpert, D. (1999). Linearly combining density estimators via stacking. *Machine Learning*, 36(1), 59-83.
- Wei, Y., & McNicholas, P. D. (2015). Mixture model averaging for clustering. *Advances in Data Analysis and Classification*, 9(2), 197-217.